REMARKS

In the Office Action dated January 24, 2007, pending Claims 1-19 were rejected and the rejection made final. Claims 1, 10, and 19, are independent claims; the remaining claims are dependent claims. On July 24, 2007, Applicants submitted a Request for Continued Examination together with an Amendment. This Amendment has not yet been acted upon by the Examiner.

Applicants and the undersigned are most grateful for the time and effort accorded the instant application by the Examiner. On October 3, 2007, and October 5, 2007, Applicants' representative conducted telephone interviews with the Examiner during which the claims of the instant application and the applied art were discussed. It was agreed that the Examiner would reconsider the Section 112 rejections and Applicants would submit a Supplemental Amendment to rewrite the claims to further elucidate the present invention in view of the Section 103 rejections.

Applicants are not conceding in this application the claims amended herein are not patentable over the art cited by the Examiner, as the present claim amendments are only for facilitating expeditious prosecution. Applicants respectfully reserve the right to pursue these and other claims in one or more continuations and/or divisional patent applications. Applicants specifically state no amendment to any claim herein should be construed as a disclaimer of any interest in or right to an equivalent of any element or feature of the amended claim.

Rejections under 35 U.S.C. §112

Claims 1, 10, and 19 stand rejected under 35 USC § 112, first paragraph, as failing to comply with the written description requirement. Specifically, the claims contain subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the art that the inventors at the time the application was filed, has possession of the claimed invention. Applicants respectfully disagree that the latest Amendments to the Claims fail to comply with the written description requirement.

In order to clarify some of the complexities of the current invention, Applicants provide the following brief summary. Referring to figure 1 (and accompanying text), the instant invention provides for "automatically extracting new words from a large amount of corpus." *Specification*, page 3, lines 11-12. The system includes:

> a section [] for segmenting a cleaned corpus with any segmentation methods...into unit sequences to form a segmented corpus; a GAST section for constructing a GAST with the unit sequences as inputs and getting counts of sub strings of these unit sequences appearing in the segmented corpus; and a section [] for filtering out false candidates before outputting true-new words.

Id at page 3, line 13-page 4, line1. The system outlined above may be applied to *any language* that can be segmented "with any segmentation method[]", *including but not limited to Chinese and Japanese. Id* at page 3, line 14. The specification focuses on Chinese and Japanese languages because "for some languages...there is no word boundary...and words are not well defined." *Id* at page 1, line 18-page 2, line 1. Because "it is very critical to define good SBs [segmentation boundaries] to split the long strings into small pieces while not losing good potential new words", it is important to clarify how boundaries may be chosen absent some clear indication (e.g. as for words in the Chinese language). *Id* at page 6, lines 12-14.

Since these SBs are critical, the specification goes on to explain how they may be chosen (for any language, including but not limited to Chinese and Japanese) by stating: "Some SB Patterns (SBPs) definitions follow: SBP A: Punctuation[] [is a] natural SB[]. SBP B: Arabic digits and alphabetic strings within the corpus are another kind of SBs. For further SBPs, we think of 2 cases." *Id* at page 6, lines 14-18. At this point, the specification goes into some detail about how alternative SPBs may be chosen for situations in which a language in encountered and neither SBP A nor SBP B are very helpful (i.e. because they don't exist). This does not mean, however, that the invention only works with such languages. To the contrary, the invention works well with languages, such as English, where SBP A or SBP B are available, because "it is very critical to define good SBs to split the long strings into small pieces while not losing good potential new words.". *Id* at page 6, lines 12-14.

Therefore, Applicants submit that the amendments to the claims previously presented (i.e. "wherein the segmenting and the splitting is not dependent upon word boundaries") are simply another way of stating that the SBs are critical, but it is not critical that the invention only be used where word boundaries are present. The invention is equally capable of handling languages such as Chinese where word boundaries are not present. Thus, Applicants respectfully request reconsideration and withdrawal of this rejection.

Claims 1, 10 and 19 stand rejected under 35 USC § 112, second paragraph. Applicant respectfully requests reconsideration and withdrawal of these rejections. It should be abundantly clear that, as stated in the previously submitted amendments and in the original specification, and as stated again above, the invention is not limited to the Chinese and Japanese languages. Thus, the claims do not contain contradictory language.

Rejections under 35 U.S.C. §103

Claims 1-3, 6-12, and 15-19 stand rejected under 35 USC § 103(a) as being unpatentable over Wang et al. (hereinafter "Wang") in view of Razin et al. (hereinafter "Razin") and further in view of Yang et al. (hereinafter "Yang"). Reconsideration and withdrawal of the present rejections are hereby respectfully requested.

The present invention is directed to a method and apparatus for automatically extracting new words from a cleaned corpus, where the corpus can be in any language that may or not have word boundaries (ranging from English or Latin to Chinese or Japanese). The instant invention segments a cleaned corpus to form a segmented corpus, splits the segmented corpus to form sub strings, and counts the occurrences of each sub strings appearing in the given corpus. Finally, the present invention filters out false candidates to output new words.

The previously submitted comments regarding Wang remain equally applicable here. As an initial matter, Wang is not directed to identifying new words. Furthermore, the portion of Wang cited by the Examiner does not deal with "splitting" the corpus; it merely deals with "segmenting" the corpus. *Wang*, col. 1, lines 45-60. Specifically, Applicants point the Examiner's attention to *Wang*, col. 1, lines 50-65, wherein the traditional tri-gram model is described, in which it should be noted that step b is segmenting and step c is predicting; *a "splitting" step is noticeably absent.* Step a *is not* splitting the string into sub-strings as contemplated by the instantly claimed invention; rather, step a (dissecting) is rather simply realizing that the corpus must be represented in some way (e.g. the individual characters, letters, numbers...). *Id.* As such, Wang does nothing to reduce the space required by a tree constructed from a large corpus other than stating that "data structure memory manager [] facilitates storage of a [] tree data structure [] across main memory...into an extended memory space, e.g. disk files on a mass storage device such as hard drive [] of computer system []." *Wang*, col. 8, lines 54-58. Thus, no approach to handling the extra memory is provided other than providing extra memory.

This stands in stark contrast to the instant invention, wherein the size problem is solved. The instant specification states that "[e]ven [if] GAST is [a] good data structure that compactly represents strings, there are practical issues to use if for ANWE. The space required [is] too large for constructing an efficient/feasible GAST from a large [corpus]." *Specification*, page 5, lines 14-16. In order to accommodate this problem, the instant invention indicates that "string[s] can be split into k equal pieces...[for example i]f a 20-character string is split into 4 equal sub strings, the saved nodes are 150." *Id* at page 6, lines 8-11. Thus, Wang does not disclose "splitting the segmented corpus to form sub strings, and counting the occurrences of each sub strings appearing in the corpus." Claim 1. Thus, not only is Wang deficient in many ways, Wang is not even aimed at solving a similar problem (i.e. finding new words and reducing the memory requirements necessary to find the new words). One skilled in the art would not even be motivated to consult or modify Wang to achieve the instantly claimed invention.

Razin fails to overcome the deficiencies of Wang as set forth above. To the extent that anything further needs to be said regarding Razin, as best understood, Razin appears to be directed to standardizing phrasing in a document; that is, not the same aim as the instantly claimed invention (e.g. finding new words). *Razin*, Abstract. Razin identifies phrases in a document to create a preliminary list of phrases, then filters and refines those phrases to create a final list of standard phrases. Razin then identifies phrase of a document that are similar to standard phrases, decides if the candidate phrase is similar enough to the standard phrase and compute phrase substitutions to determine the approximate conformation of the standard phrase to the approximate phrase and vice versa. *Id.* There is no suggestion or teaching in Razin that the segmenting and the splitting of the corpus is not dependent upon word boundaries. In fact, Razin teaches away from this ability (col. 11, lines 14-36), teaching that the source text is segmented using a standard finite-state machine technique that recognizes patterns that indicate word and sentence boundaries. Further, there is no suggestion or teaching that Razin discloses determining new words based upon the domain of the current corpus.

As best understood, Yang appears to be directed towards Chinese language modeling. Yang fails to overcome the deficiencies of Wang and Razin as asserted above. Specifically, Yang fails to teach, *inter alia*, "splitting" of the strings into sub-strings, as recited in the independent claims. Claim 1. As noted in the original specification, it is of no consequence that some languages do not have clear word boundaries (e.g. Chinese) because "any segmentation methods, such as…statistic segmentation which [is] used widely" can be employed to segment the corpus. *Specification,* page 3, lines 14-15. Thus Yang takes nothing away from the instantly claimed invention, and certainly does not render it obvious either alone or in any combination with Wang or Razin. Thus, Applicants respectfully request reconsideration and withdrawal of these claim rejections.

Claims 4-5 and 13-14 stand rejected under 35 USC § 103(a) as being unpatentable over Wang et in view of Razin and Yang and further in view of Hui. Reconsideration and withdrawal of this rejection is hereby respectfully requested.

Hui cannot overcome the deficiencies of Wang, Razin and Yang as discussed above. Hui is incorporated by reference in the original specification to give an overview of the AST construction and GAST. Nothing in Hui renders the instantly claimed invention obvious, either alone or in any combination with Wang, Razin or Yang. Thus, reconsideration and withdrawal of these rejections is respectfully requested.

Solely in an effort to expedite prosecution of the instant application, the independent claims have been amended to recite, *inter alia*, "wherein the new words are words not contained in a base vocabulary." Claim 1. Applicants respectfully submit that the definition of "new words" was already clear from the original specification and the original claim language. As can be readily understood from the specification, wherein it states "since [example compound character word] is a known word from the common base vocabulary, it can be omitted", new words are not known or "old" words. Nonetheless, Applicants respectfully submit that there can be no misunderstanding as to the meaning of "new words" in the claims due to these amendments.

In view of the foregoing, it is respectfully submitted that independent Claims 1, 10 and 19 fully distinguish over the applied art and are thus allowable. By virtue of dependence from Claims 1 and 10, it is thus also submitted that Claims 2-9 and 11-18 are also allowable at this juncture.

In summary, it is respectfully submitted that the instant application, including Claims 1-19, is presently in condition for allowance. Notice to the effect is hereby earnestly solicited.

Respectfully submitted,

Stanley D. Ference III
Registration No. 33,879

Customer No. 35195
FERENCE & ASSOCIATES LLC
409 Broad Street
Pittsburgh, Pennsylvania 15143
(412) 741-8400
(412) 741-9292 - Facsimile

Attorneys for Applicants